Application of a Propensity Score Approach for Risk Adjustment in Profiling Multiple Physician Groups on Asthma Care

I-Chan Huang, Constantine Frangakis, Francesca Dominici, Gregory B. Diette, and Albert W. Wu

Objectives. To develop a propensity score-based risk adjustment method to estimate the performance of 20 physician groups and to compare performance rankings using our method to a standard hierarchical regression-based risk adjustment method.

Data Sources/Study Setting. Mailed survey of patients from 20 California physician groups between July 1998 and February 1999.

Study Design. A cross-sectional analysis of physician group performance using patient satisfaction with asthma care. We compared the performance of the 20 physician groups using a novel propensity score-based risk adjustment method. More specifically, by using a multinomial logistic regression model we estimated for each patient the propensity scores, or probabilities, of having been treated by each of the 20 physician groups. To adjust for different distributions of characteristics across groups, patients cared for by a given group were first stratified into five strata based on their propensity of being in that group. Then, strata-specific performance was combined across the five strata. We compared our propensity score method to hierarchical model-based risk adjustment without using propensity scores. The impact of different risk-adjustment methods on performance was measured in terms of percentage changes in absolute and quintile ranking (AR, QR), and weighted κ of agreement on QR.

Results. The propensity score-based risk adjustment method balanced the distributions of all covariates among the 20 physician groups, providing evidence for validity. The propensity score-based method and the hierarchical model-based method without propensity scores provided substantially different rankings (75 percent of groups differed in AR, 50 percent differed in QR, weighted κ = 0.69).

Conclusions. We developed and tested a propensity score method for profiling multiple physician groups. We found that our method could balance the distributions of covariates across groups and yielded substantially different profiles compared with conventional methods. Propensity score-based risk adjustment should be considered in studies examining quality comparisons.

Key Words. Physician group, profiling, propensity score, regression-to-the-mean, risk adjustment

Provider profiles are used increasingly to compare performance, increase provider accountability, help health care managers to monitor quality of care, and help consumers to choose providers or health plans (Enthoven 1993; Bodenheimer 1999). However, comparisons of provider performance can be biased when patients cared for by different providers differ in background characteristics. Without appropriate risk adjustment, providers who care for sicker patients may appear to perform worse, and patients may be misled about the relative quality of care.

For quality assessment, random assignment of patients to different health care providers would be ideal to balance the distributions of patient characteristics among providers, thus removing confounding. However, it is neither practical nor desirable to randomly assign patients to different providers, because, for example, patients with a specific condition may gravitate to certain providers who specialize in such care. What are required are methods for risk adjustment that are valid in the context of nonrandom selection. In observational studies, statistical risk-adjustment techniques are used to remove confounding effects (Iezzoni 1997). The most common method for risk adjustment is regression modeling (DeLong et al. 1997; Shahian et al. 2001). However, the standard regression-based risk adjustment is limited because it does not ensure balance in the distributions of covariates among providers (Dehejia and Wahba 1999). The importance of balancing increases with the number of covariates (Rubin 1997).

The propensity score was originally proposed as a method for producing balance of many covariates between two groups (Rosenbaum and Rubin 1983, 1984). This method can balance a set of many covariates by estimating the probability (propensity) of assignment to a specific provider given those

Conflicts of interests: There are no conflicts of interests for any authors. Presented in part at the following conferences: the 19th Annual Research Meeting of AcademyHealth, June 23–25, 2002, Washington DC, and the 2003 Joint Statistical Meeting (JSM), August 3–7, 2003, San Francisco.

Address Correspondence to Albert W. Wu, M.D., M.P.H, Department of Health Policy and Management, Bloomberg School of Public Health, The Johns Hopkins University, 624 North Broadway/Room 633, Baltimore, MD 21205-1901. I-Chan Huang, Ph.D., is with the Department of Health Policy and Management, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore. Constantine Frangakis, Ph.D. and Francesca Dominici, Ph.D. are with the Department of Biostatistics, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore. Gregory B. Diette, M.D., M.H.S. and Albert W. Wu, M.D., M.P.H. are with the Department of Epidemiology, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore and also with Department of Medicine, School of Medicine, The Johns Hopkins University, Baltimore.

covariates. For observed covariates, theory assures that given any value of the propensity score, the subgroups of patients who enroll with different providers will have the same joint distribution in all the covariates that were used to estimate that propensity score (Rosenbaum and Rubin 1983, 1984; Rubin 1997). This is a main advantage of propensity score methods, because it allows a straightforward check for whether the adjustment has made providers comparable with respect to the observed covariates (Rosenbaum and Rubin 1983, 1984; Rubin 1997). Propensity score-based risk adjustment also assures that if enrollment to different providers is "ignorable" based on the observed covariates (i.e., enrollment is not affected by unobserved patient or provider characteristics) (Rosenbaum and Rubin 1983), then enrollment is also ignorable given only the propensity score.

In practice, there can be direct or indirect evidence that the propensity score is better than standard methods for estimating the true underlying difference of comparison groups. Direct evidence exists only when the study is controlled. For example, in a randomized study, Lalonde compared the effect of a training program designed to help disadvantaged workers increase earnings (LaLonde 1986). In this study, evaluation using standard regression models could not replicate the experimentally determined results. However, using the same data set, propensity score techniques produced results similar to those of the randomized experiment (Dehejia and Wahba 1999).

Without a controlled design, the true unconfounded differences are not known, and indirect evidence is used to judge suitability of the propensity score method. Indirect evidence exists when (1) the propensity score method has balanced all important observed covariates between the comparison groups; and (2) the results from the propensity score method differ from those when not using propensity scores.

Propensity score techniques were originally designed for two-group comparisons (Rosenbaum and Rubin 1983, 1984), and have been used in observational studies with cohort or case-control designs to reduce bias from estimated effects of treatment programs (Connors et al. 1996; Shwartz et al. 1999; Gum et al. 2001), and social (Dehejia and Wahba 1999) or health services programs (Keating et al. 2001; Mojtabai and Zivin 2003). Imbens (2000) developed a modified method for comparison of multiple groups. To our knowledge, such a method has not been used in health services research for profiling multiple providers. In addition, with multiple providers, provider-specific estimates of performance are subject to regression-to-the-mean because of small numbers within provider (Christiansen and Morris 1997); this issue has not been addressed using propensity scores.

The goals of this study were (1) to develop and validate a propensity score-based risk adjustment method to estimate performance of multiple providers, in order to balance all observed covariates, as well as to address regression-to-the-mean; and (2) to compare this method versus a more conventional outcome regression method of evaluating and ranking performance in 20 California physician groups. Satisfaction with asthma care was used as the performance indicator. The outcome regression-based method adopted in this study is a hierarchical model that adjusts for the regression-to-the-mean, but without using the propensity score (Morris 1983; Christiansen and Morris 1997; Sullivan, Dukes, and Losina 1999).

We hypothesized that the propensity score-based method would balance all observed covariates. If, in addition, the propensity score method also results in substantial ranking differences in physician group performance compared with the standard method, this finding will provide indirect evidence for greater usefulness of the propensity score method.

Methods

Study Sample

This study was conducted in conjunction with 20 California physician groups that participated in the 1998 Asthma Outcomes Survey (AOS). The AOS was initiated by the Pacific Business Group on Health (PBGH) a health care purchasing coalition in California, and HealthNet, a California-based health plan, to evaluate, improve, and report on the quality of asthma care at the physician group level (Masland et al. 2000). The 20 participating physician groups were instructed to use administrative materials to identify all managed care patients with at least one asthma-related encounter in the outpatient, emergency, or inpatient settings (identified by ICD-9 code 493.xx) between January 1, 1997 and December 31, 1997. Patients had to be continuously enrolled in the physician group for that calendar year. From eligible patients, this study randomly selected a sample of 650 patients from each physician group. If a physician group had fewer than 650 eligible patients, then all eligible patients were sampled. A total of 7,820 patients had usable addresses and met the study eligibility criteria.

Data Collection

Patient data were collected by self-administered mailed survey. The survey instrument was largely based on the "Health Survey for Asthma Patients" developed at the Johns Hopkins Health Services Research and Development Center for the Outcomes Management System Consortium Asthma Project of the Managed Health Care Association (Steinwachs, Wu, and Skinner 1994; Diette et al. 1999). The instrument used in this study included questions relating to patient characteristics, general health, asthma symptoms, effect of asthma on functioning, asthma medications and treatment, self-management knowledge and activities, access to care, and patient satisfaction.

The survey period began in July 1998 and ended in February 1999. The survey was fielded by PBGH and HealthNet using an identical methodology. A total of 2,515 responses were obtained for response rate of 32.2 percent, which is typical on satisfaction surveys (Fowler et al. 2002). Response rates also differed somewhat across physician groups. We did not have patient characteristics for nonrespondents, so were not able to test if, within physician group, satisfaction scores differed systematically between respondents and nonrespondents. However, the lower response rate seems unlikely to affect the comparison among regression- and propensity score-based methods.

Risk Adjustors and Performance Indicator

All of the variables used in this study, including risk adjustors and outcome indicators, were from the patient survey. Satisfaction with asthma care was selected as the performance indicator. In the survey instrument, patient satisfaction was rated on a five-point Likert-type scale (Poor/Fair/Good/Very Good/Excellent). We dichotomized this variable into "greater satisfaction (Very Good/Excellent)" versus "less satisfaction (Poor/Fair/Good)."

Analytic Framework for Comparing Different Risk-Adjustment Methods

We adjusted for exogenous factors, i.e., factors for which providers have no influence (mainly patient characteristics, such as age, sex, education, baseline severity, etc.). We did not include race in the risk-adjustment model. Evidence suggests that African-American patients may receive poorer quality of care than white patients (Kahn et al. 1994). If this makes patient satisfaction across physician groups differ because of different race distribution, these are differences we want to capture, and adjusting for race here would mask them. In this case, examining patient satisfaction separately within race groups would highlight such inequalities (Iezzoni 1997). Moreover, we did not adjust for endogenous factors, i.e., factors that providers can affect (mainly physician group characteristics, such as physician group specialty, number of supplementary staff, etc.) (Welch, Black, and Fisher 1995). Adjusting for endogenous factors may mask true performance of physician groups because these factors can influence the patient outcomes.

We evaluated patient satisfaction among physician groups using two analytic methods, thus also assessing sensitivity of the results to different risk-adjustment approaches (Table 1). We used the first physician group as the reference group for comparisons among different methods.

For method 1, we implemented a hierarchical outcome regression model without propensity scores. At the first stage (patient level) we used a logistic regression model for estimating the group-specific log odds ratio (OR) of patient satisfaction (outcome) as a function of patient characteristics, including age, sex, education level, type of insurance, prescription drug coverage, asthma severity, number of comorbidities, and health status. At the second stage (group level), we modeled the variation of the log OR across 20 physician

Table 1: Analytic Framework for Comparing Risk-Adjustment Methods for Physician Group Profiling

		Risk-Adjustment Method	
Method	*Description*	*Risk Adjustor*	*Remarks*
Method 1	Hierarchical outcome regression adjustment without propensity scores	Sociodemographic (age, sex, education level, types of insurance, drug coverage), Clinical (asthma severity and number of co-morbid conditions), Health status (SF-36 physical and mental component scores)	1. Adjusts for covariate effects on patient satisfaction 2. Addresses regression-to-the-mean using hierarchical regression on the covariates
Method 2	Propensity score-based risk adjustment	Same as for Method 1	1. Adjusts for covariate effects on provider selection, using propensity scores; does not adjust for effects on satisfaction 2. Addresses regression-to-the-mean using shrinkage techniques* on the propensity-score based proportions of satisfaction

*Using Morris's approach (Morris 1983).

groups. The hierarchical outcome regression approach takes into account clustering of patients within physician groups and the different number of patients within each physician group (reliability). Under the hierarchical outcome model the group-specific estimates of performance are shrunk toward an average performance common to all physician groups to address the regression-to-the mean that arises with comparison of multiple groups (Morris 1983; Christiansen and Morris 1997; Sullivan, Dukes, and Losina 1999). The hierarchical outcome regression model is detailed in Appendix 1.

With this method, the relative performance of physician groups was assessed by estimating the risk-adjusted OR of satisfaction with care (greater versus less satisfaction) attributable to the jth physician group relative to the first physician group (reference group) by exponentiating the difference between the estimated provider-specific random intercept of the jth ($j = 2, \ldots, 20$) and the first physician group (DeLong et al. 1997; Katon et al. 2000).

For method 2, we implemented a propensity score-based risk adjustment. With this method, the main goal was to estimate the proportions, p_j, of satisfied patients in the hypothetical scenario under which all patients would have been enrolled in the jth group ($j = 1, \ldots, 20$). Then, performance was compared among physician groups by first obtaining preliminary estimates, P_j^+, of the proportions p_j ($j = 1, \ldots, 20$). We obtained P_j^+ by adapting Imbens's propensity score method for multiple groups (Imbens 2000), and then obtained final estimates of p_j with a method that accounts for regression-to-the-mean. Specifically, we developed five major steps for the propensity score method for the multiple physician groups: (1) calculation, for each patient, of 20 estimated propensity scores, each being the probability of enrollment in a particular physician group versus all the remaining groups, (2) stratification of patients into quintiles for each of the 20 physician groups based on the propensity scores, (3) validation of estimation of the propensity scores (also called a balance check), (4) estimation of the preliminary adjusted proportion of satisfaction, P_j^+, of each physician group by combining across the five propensity strata, and (5) estimation of the relative performance of each physician group using shrinkage techniques. Note that in step 2, the same patient's stratification may be different for different physician groups, since the propensity scores can change. These five steps are detailed in the Appendix 2.

Comparison of Impact of Different Risk-Adjustment Methods on Profiling Rankings

There is no consensus on how to quantify ranking impact (or ranking change) on provider performance. However, rank-based measures are popular in the practice of comparing provider profiling (Aron et al. 1998; DiGiuseppe et al. 2001). In this study, the impact of different risk-adjustment methods on physician group profiling was measured in terms of differences in estimated performance ranking between the two methods. Rankings of physician groups were compared based on the OR of greater satisfaction versus less satisfaction for the *j*th physician group versus the reference group.

Two methods were used to demonstrate changes in ranking: percentage changes in absolute ranking (AR) and percentage changes in quintile ranking (QR). In practice, the percentage changes in QR are more useful for consumer choice or performance reward than the percentage changes in AR (Alter et al. 2002). Percentage changes in AR represented the portion of physician groups that changed in ranking. The QR represented the portion of physician groups that moved into a different quintile of ranking. It was evaluated using a weighted κ statistic to measure the significance of these differences in ranking to those expected by chance. We used quadratic-weighted κ rather than standard κ (no weight) to reflect the ordinal nature (quintile) of the ranking scale (Streiner and Norman 1995). To date, there is no standard to judge the AR and QR. For QR being measured using weighted κ statistics, it is usually recommended to use 0.7 as the cutoff for acceptable agreement. However, others have recommended using 0.5 as the criterion (Fayers and Machin 2000).

Statistical Package

We used *SAS* 8.1 with Glimmix Macro for hierarchical modeling analysis. For the propensity score, we used *STATA* 7.0 combined with existing routines for shrinkage (Everson and Morris 1993) (see http://www.biostat.jhsph.edu/~cfrangak/papers/proscore_profiling).

Results

Characteristics of Physician Groups and Respondents

Of the 20 participating physician groups, eight were located in Northern and 12 in Southern California. The characteristics of the 2,515 participants are shown in Table 2. Patients ranged in age from 18 to 56 years with a mean age of 39.9 years (SD: 9.5); 71.2 percent were female, 70.3 percent were white, and 5.1 percent were African American; 81.6 percent had at least some college education; 69.1 percent obtained health insurance through their employer, and 24.8 percent by themselves; and 96.5 percent had prescription drug

Table 2: Characteristics of Patients with Asthma ($n = 2{,}515$)

Dimension	*Frequency or Mean (SD)*
Age (%)	
18–24	7.2
25–34	22.0
35–44	34.6
45–54	33.2
55 and above	3.1
Overall, mean (SD)	39.9 (9.5)
Sex (%)	
Male	28.8
Female	71.2
Race (%)	
White	70.3
African American	5.1
Asian American	10.0
Other	14.7
Education (%)	
High school or below	18.4
College	65.3
Graduate	16.3
Health insurance status (%)	
Private—through employer	69.1
Private—through self-purchase	24.8
Public—Medicare, Medicaid	1.4
Other	4.9
Drug insurance coverage (%)	96.5
Asthma severity (%)	
Mild intermittent	14.4
Mild persistent	19.2
Moderate persistent	49.3
Severe persistent	17.1
Number of comorbidity, mean (SD)	2.1 (1.4)
SF-36 Physical component score, mean (SD)	45.7 (10.3)
SF-36 Mental component score, mean (SD)	47.4 (10.7)
Satisfaction with asthma care	
More satisfied with asthma care	55.4
Less satisfied with asthma care	44.7

coverage. On clinical characteristics, 14.4 percent had mild intermittent asthma, 19.2 percent had mild persistent asthma, 49.3 percent had moderate persistent asthma, and 17.1 percent had severe persistent asthma. The mean number of comorbidities was 2.1 (SD: 1.4). For general health status, the mean SF-36 physical component score (PCS) was 45.7 (SD: 10.3), and the mean SF-36 mental component score (MCS) was 47.4 (SD: 10.7). The SF-36 PCS and MCS have been standardized to the U.S. general population (mean score of

50 ± 10) to allow easier norm-based interpretations. The higher SF-36 PCS and MCS scores indicate better health states (Ware 1997).

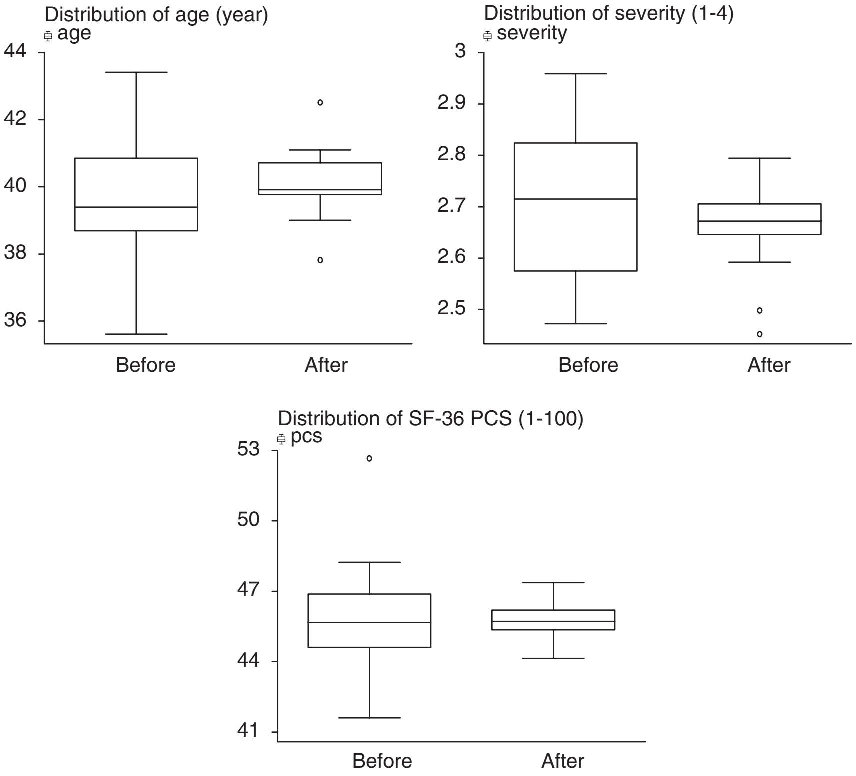
Balancing Covariates across Multiple Physician Groups Using Propensity Scores

Before applying the propensity score method, there was imbalance in each covariate across 20 physician groups. For example, the range of mean patient age among the 20 physician groups was 35.6–43.4 (SD: 1.9) ($p<.01$); the range of mean severity was 2.5–3.0 (SD: 0.49) ($p<.01$); the range of mean SF-36 PCS was 41.6–52.7 (SD: 2.19) ($p<.01$). The distributions of sex, level of education, type of health insurance, prescription drug coverage, and number of comorbidities were also significantly unbalanced across the 20 physician groups (all $p<.01$). The difference in distribution of SF-36 MCS was marginally significant ($p=.05$).

After applying the propensity score method, the balance of each covariate across the 20 physician groups improved substantially. Figure 1 shows the ranges of the group-specific averages of the important covariates of age, asthma severity, and SF-36 PCS across the 20 physician groups before adjustment (left part of each graph) and after adjustment (right part of each graph) using the propensity score techniques, as a result of step 3. Specifically, for covariate X, we define the adjusted average $\tilde{X}_j$ for physician group j as $\tilde{X}_j = \sum_s \left(\bar{X}_{j,s} * W_{j,s}\right)$, where $\bar{X}_{j,s}$ is the average of the covariate in the estimated propensity stratum s of physician group j, and $W_{j,s}$ are the weights given in step 4 of Appendix 2. If the estimated propensity score equals the true propensity score, and not generally in other situations, then each adjusted average $\tilde{X}_j$ approximates a common average, the average of the covariate in the population (the proof is the same as that for the potential outcomes in Rosenbaum and Rubin [1983]). Therefore, similarity among the plotted $\tilde{X}_j$, $j=1,\ldots,20$, compared with among the unadjusted averages, as shown in the right and left parts, respectively, of the graphs in Figure 1, also graphically supports how well one has approximated the propensity score.

After propensity score adjustment, the standard deviation for age was reduced from 1.9 to 0.93 (51.1 percent reduction) and the range was reduced from 7.8 to 4.7. For asthma severity, the standard deviation was reduced from 0.14 to 0.08 (42.9 percent reduction) and the range was reduced from 0.49 to 0.34. For SF-36 PCS, the standard deviation was reduced from 2.19 to 0.67 (69.4 percent reduction) and the range was reduced from 11.04 to 3.23. For the other covariates, the ranges of distributions were also significantly reduced. After adjustment, only 3.9 percent of all comparisons for balance status (7 out

Figure 1: Distributions of Group-Specific Averages of Patient Age, Asthma Severity, and SF-36 PCS among the 20 Physician Groups before Adjustment (Numbers Are Plain Averages within Physician Groups) and after Adjustment (Numbers Are Described in Paragraph 3 of Results) with the Propensity Scores



of 180 comparisons) were statistically significant at the level of $p<.05$, indicating that the propensity score adjustment produced balance, in the observed covariates, similar to that which would be expected by randomization of these covariates across physician groups.

Comparison of Rankings Impact on Physician Group Performance Based on Different Risk-Adjustment Methods

Table 3 shows the unadjusted and adjusted performance of 20 physician groups based on different methods. When comparing the propensity score-based method (method 2) with the hierarchical model-based method (method

Table 3: Performance of 20 Physician Groups Estimated Using Different Methods

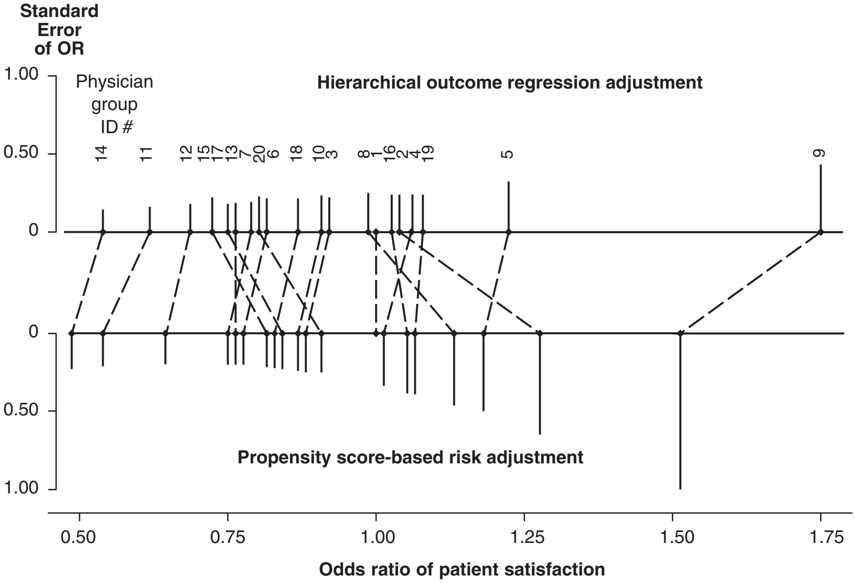
		No Risk Adjustment		*Hierarchical Outcome Regression Adjustment*		*Propensity Score-Based Risk Adjustment*	
Group ID	*Number of Patients in Group*	*Unadjusted Rate (%)*	*OR (SE)*	*Adjusted Rates (%)*	*OR (SE)*	*Adjusted Rates (%)*	*OR (SE)*
1*	163	63.8	1.0	64.7	1.0	57.7	1.0
2	177	60.5	0.87 (0.19)	65.4	1.03 (0.23)	63.7	1.29 (0.68)
3	151	58.3	0.79 (0.18)	62.2	0.90 (0.21)	54.5	0.88 (0.24)
4	212	59.0	0.82 (0.17)	65.8	1.05 (0.23)	57.9	1.01 (0.34)
5	63	71.4	1.42 (0.46)	68.9	1.21 (0.32)	61.6	1.18 (0.52)
6	86	59.3	0.83 (0.23)	59.4	0.80 (0.21)	51.2	0.77 (0.20)
7	146	49.3	0.55 (0.13)	58.5	0.77 (0.18)	50.5	0.75 (0.19)
8	82	58.5	0.80 (0.22)	64.2	0.98 (0.25)	60.8	1.14 (0.48)
9	110	78.2	2.03 (0.57)	76.4	1.77 (0.45)	67.7	1.54 (1.05)
10	75	53.3	0.65 (0.18)	62.0	0.89 (0.23)	54.2	0.87 (0.23)
11	64	37.5	0.34 (0.10)	52.8	0.61 (0.17)	42.3	0.54 (0.20)
12	103	47.6	0.51 (0.13)	55.1	0.67 (0.17)	46.1	0.63 (0.19)
13	176	48.9	0.54 (0.12)	57.9	0.75 (0.17)	50.8	0.76 (0.20)
14	141	36.9	0.33 (0.08)	49.2	0.53 (0.13)	39.0	0.47 (0.22)
15	31	38.7	0.36 (0.14)	56.5	0.71 (0.21)	52.4	0.81 (0.21)
16	164	61.6	0.91 (0.21)	65.1	1.02 (0.24)	59.0	1.06 (0.40)
17	194	48.5	0.53 (0.12)	57.5	0.74 (0.17)	53.6	0.85 (0.22)
18	110	50.9	0.59 (0.15)	60.9	0.85 (0.21)	52.7	0.82 (0.21)
19	218	58.7	0.81 (0.17)	66.0	1.06 (0.23)	59.3	1.07 (0.40)
20	49	49.0	0.54 (0.18)	59.1	0.79 (0.22)	55.0	0.90 (0.24)

*Physician group 1 as the reference group.
OR = odds ratio.

1), there was a 75 percent difference in AR and 50 percent difference in QR, with a weighted κ of 0.69.

Figure 2 shows the correspondence in ORs of patient satisfaction (in the *j*th physician groups relative to the reference group) as estimated by using the hierarchical outcome regression without the propensity score (method 1) versus the propensity score-based risk adjustment (method 2). The figure is especially useful for characterizing where the differences between the two methods occur. In general, differences in rankings fell into two clusters. For ARs, most of the shifts occurred within the 80 percent middle ranks of physician groups. For QRs, five physician groups (ID numbers 4, 6, 7, 18, and 19) shifted their QR into a lower quintile after propensity score adjustment. Also, five physician groups (ID numbers 2, 8, 15, 17, and 20) shifted their QR into a higher quintile. Table 4 shows that those providers that changed QR from

Figure 2: Ranking Shift between Hierarchical Outcome Regression Adjustment and Propensity Score-Based Risk Adjustment



hierarchical outcome regression method to propensity score-based method usually have some distinct characteristics. For example, physician group 8, which increased in rank from 8 to 4 is characterized by younger patients, more females, poor asthma severity, more comorbid conditions, less prescription drug coverage, and poor physical as well as mental health. However, physician groups with the best (ID number 9) or worst (ID numbers 11, 12, and 14) performance in the hierarchical outcome regression method did not shift in ranking after applying the propensity score.

Compared with the hierarchical outcome regression method, the propensity score-based method produced only slightly larger standard errors of the ORs. This is because the estimates from the outcome regression method are theoretically efficient if the model is correct, but its standard errors do not increase if the model is incorrect.

DISCUSSION

To accurately compare provider performance, it is critical to control for differences in the characteristics of patients treated by different providers. Owing

Table 4: Characteristics of Physician Groups That Shifted Quintile Rankings Based on Different Risk Adjustment Methods*

ID # of Physician Group	*Location*	*Number of Patients in Group*	*Patient Characteristics*[†]
(A) Raising ranks after using propensity score method			
2	Northern California	177	Gender, severity, number of comorbidity, drug prescription coverage, PCS, MCS
8	Northern California	82	Age, gender, severity, number of comorbidity, drug prescription coverage, PCS, MCS
15	Southern California	31	Age, gender, severity, number of comorbidity, drug prescription coverage, PCS, MCS
17	Southern California	194	Age, gender, PCS, MCS
20	Southern California	49	Age, gender, severity, number of comorbidity, PCS, MCS
(B) Lowering ranks after using propensity score method			
4	Northen California	212	Age, gender, severity, number of comorbidity, PCS
6	Northen California	86	Age, gender, severity, number of comorbidity, drug prescription coverage, PCS, MCS
7	Northern California	146	Age, gender, number of comorbidity, drug prescription coverage
18	Southern California	110	Age, gender, severity, number of comorbidity, drug prescription coverage, PCS
19	Southern California	218	Age, gender, drug prescription coverage, PCS, MCS

*Method 1 (hierarchical outcome regression adjustment without using propensity score), Method 2 (propensity score-based risk adjustment).

[†]Statistically different from grand mean ($p<0.05$).

PCS = physical component score; MCS = mental component score.

to the difficulty of designing a randomized experiment to compare provider group performance, risk adjustment is used to account for background differences. In this study, we applied a novel propensity score method to compare the performance of multiple physician groups.

Our results showed that the propensity score method improved the balance of covariates among physician groups, leaving imbalance similar to that which would be expected by randomization of these covariates. Moreover, the propensity score-based method produced ranking results that differed from those using the outcome regression method, suggesting that profiling is sensitive to patient selection bias whether or not it is controlled. In the absence of a controlled experiment, we cannot provide direct evidence that the propensity score-based method is superior to the outcome regression method. However, the above two results taken together do provide indirect evidence that the propensity score method is more reliable than the outcome regression without using the propensity score, because the latter method cannot ensure comparability of the distribution of patient covariates across providers and yields different results from the propensity score method that can assure such comparability.

From a methodological point of view, there are several advantages of applying the propensity score method. First, it allows researchers to compare provider performance with similar patient characteristics without specification (or assumption) of a linear relationship between the profiling indicator and risk adjustors as is required by a regression-based method. Some risk adjustors, such as patient's age, may not fit this linear assumption (Iezzoni 1997). The application of regression-based risk adjustment to balance the distributions of covariates across providers is particularly limited when patient characteristics are more skewed for some providers than others, since this involves extrapolation where there is little overlap of the covariate distributions (Rosenbaum and Rubin 1983). Propensity score methods model the assignment of patients (rather than the outcome or performance indicators) to a specific provider based on patient characteristics. Although propensity scores also involve modeling, there is a simple model-checking requirement, which is the balance of covariates (see Introduction). Therefore, in practice, propensity score methods can be more robust to model misspecification than regression-based methods.

It should be noted that the propensity score methods might not perfectly balance the distribution of covariates across providers because of estimation error and a suboptimal matching algorithm. In general, there are four ways to use propensity scoring: matching, regression, propensity score weighting (or inverse probability weighting [IPW]), and stratification (Rosenbaum and Rubin 1983; D'Agostino 1998). Propensity matching has gained popularity with empirical researchers and theorists; however, estimates are sensitive to the choice of matching algorithms. Some algorithms could induce bias (Rosenbaum and Rubin 1985), so other matching algorithms are considered optimal (Rosenbaum 1989). For propensity regression, evidence showed that using propensity score as a regressor could be biased even with higher order items (Dehejia and Wahba 1999). Although the application of propensity score weighting to profiling has not been published in the medical literature, econometricians found that propensity score weighting could achieve the efficiency of the maximum likelihood estimate (Hirano and Imbens 2002). The difference between the stratification method and the method that weighs each participant exactly on the estimated propensity score is more practical than conceptual. For propensity score weighting, each person is his/her own stratum, and there are as many strata as there are subjects, whereas the stratification method has only a few strata to reweigh. In our case, we applied the stratification method instead of IPW because (1) Cochran's pioneering work in the field of observational studies has suggested that a few, typically five, strata are enough to reduce the bias of confounding by measured variables (Cochran 1968), and (2) Dehejia and Wahba (1999) have shown that treating each person as their own stratum and weighting on the propensity score, without some coarsening or other adjustment, leads to high variance (in fact, we may get a rate larger than 100 percent) and hence low overall accuracy. For comparison, we also calculated the 20 adjusted satisfaction rates for the 20 physician groups using IPW. The results were more variable than both of the methods in this paper. Moreover, the rankings with IPW were not agreeing even for physician groups for which the other two methods (hierarchical outcome regression and propensity score) were agreeing with each other.

To date, the development of risk-adjustment methods for health services research has been limited. Most efforts have emphasized careful selection of risk adjustors (Iezzoni 1997), while few have focused on the balancing of patient selection (Christiansen and Morris 1997; DeLong et al. 1997; Sullivan, Dukes, and Losina 1999), and none has underscored the importance of the balancing of covariates. It is important to clarify that different risk-adjustment methods may be appropriate, depending on their purposes. For setting premium rates, it is desirable to develop models that can predict individual patients' future costs based on specific individuals' values of a group of risk adjustors. For that purpose, an outcome regression model is a practical prediction tool. If, however, the purpose is to compare overall provider performance, it is important to properly balance observed covariates as can be done with propensity score methods.

For practical use in provider profiling, we recommend using a propensity score-based method to refine and complement regression-based risk adjustments. A general regression-based method can be used first to select a subset of risk adjustors, followed by application of propensity-score techniques to balance those risk adjustors among providers. To identify the best and worst providers for benchmarking or quality management, it may be useful to plot ranking shifts based on different methods as demonstrated in Figure 2. For rankings for which both methods agree, we can be more confident in the results. For the rankings for which the methods do not agree, and so long as the standard errors are comparable between the two methods, balancing of the covariates resulting from the propensity score method is likely to be more trustworthy.

In interpreting our findings, several limitations should be noted. First, the set of risk adjustors included in our risk-adjustment models may not be optimal. In this study, all of the risk adjustors were collected from the patient survey. We did not collect clinical assessments, some patient characteristics (e.g. personal income or family size), and other nonpatient characteristics that providers cannot influence (e.g. health plan or physician group penetration rate), which could be confounding (Braitman and Rosenbaum 2002). Thus, our propensity score method can balance unobserved covariates only to the extent that those are correlated with the observed covariates (Rubin 1997).

Second, related to estimability is also the situation whereby the set of covariates would be different for different participants. This would mean that we would have as many models as participants, and estimation would not be feasible, unless new data and models were posited. A possible way, for example, to address this could be by using an instrumental variable. However, we could not identify a variable that would serve as a good instrument from these data. Moreover, note that such inestimability arising by allowing as many models as there are participants is a problem that could arise in most other approaches that involve many covariates.

Third, even after adjusting of the covariates with a propensity score, there is not a single way of addressing regression-to-the-mean to improve accuracy of the overall estimates. For example, it can be possible to develop and justify shrinkage methods that report different rankings of groups only if in some scale, for example using the standard errors, there is enough information to do so.

In conclusion, we propose a novel propensity score method to compare performance across multiple physician groups by properly balancing patient-specific covariates across physician groups, and by taking into account the clustered nature of the data. The paper's results support the original properties of the propensity score in this setting. However, in the absence of a relevant experiment in this area, the conclusions of our comparison between the methods are not certain. We hope that this paper instigates further work in this area.

Acknowledgments

We greatly appreciate the assistance of Cheryl L. Damberg, Ph.D. in conducting this study. We also thank the editors and two anonymous reviewers for their suggestions.

Appendix 1: Hierarchical Outcome Regression Model to Compare Multiple Physician Groups

The hierarchical outcome regression model is

$$\text{Logit } P\left(Y_{ij} = 1|X_{hij}\right) = \beta_{0j} + \sum_{h=1}^{p} \beta_h X_{hij}$$

$$\beta_{0j} = \beta_0 + \mu_{0j}; \quad \mu_{0j} \sim N\left(0, \tau_{00}\right)$$

where i is the subject index, j is the physician group index, X_{hij} is covariate h of subject i in physician group j, β_{0j} is the group-specific random intercept, β_0 is the overall mean intercept, β_h is the overall slope for subject characteristic h, μ_{0j} is the intercept random effect of physician group j, τ_{00} is the variance of group intercepts.

Estimation of the above model is done with restricted maximum likelihood for the fixed-effect parameters β_0, β_h, and τ_{00}, and empirical Bayes for the intercept random effects μ_{0j}. These estimates cannot be given in a closed form but they are analogous in spirit to Morris's shrinkage method as described in step 5 of Appendix 2. Morris's method has been recommended as superior for the case where, either covariates have been already addressed, as with the propensity score in Appendix 2, or if covariates vary only across groups, but is not applicable, at least in its original form, when covariates vary also within groups, as in the outcome regression.

Appendix 2: Propensity Score Method to Compare Multiple Physician Groups

The goal was to estimate the proportions, p_j, of satisfied patients, if all patients had been enrolled in group j ($j = 1, \ldots, 20$). There were five major steps to use propensity scores to estimate these proportions.

Step 1: Calculation of the Propensity Score, e_{ij}*, of Patient* i *Enrolling in the* j*th Physician Group* (j = *1, . . ., 20*)

Each patient has 20 propensity scores, each of them representing the probability of enrollment in each of the 20 physician groups. The propensity score e_{ij} was estimated as the conditional probability of patient i to have been enrolled in the jth physician group ($j = 1, \ldots, 20$) as a function of the patient's specific covariates (Table A1). The preliminary propensity scores estimates were obtained by using a multinomial logistic regression model.

Step 2: Stratification of Patients into Quintiles Based on the Propensity Scores

For evaluating the jth physician group the propensity scores of all 2,515 patients (the jth row of Table A1), as estimated from step 1, were ranked and then stratified into five strata based upon quintiles (Figure A1). Evidence has shown that such stratification based on the quintiles of the scores generally reduces bias because of unbalanced covariates by 90 percent (Cochran 1968; Rosenbaum and Rubin 1984).

Patients in these five strata (low to high propensity scores) were then stratified by whether or not they actually belonged to the jth group, and then further stratified by whether or not they were satisfied with asthma care (Figure A2). This three-way stratification for the jth group is also shown in Table A2.

Step 3: Validation of Propensity Scores Estimates

The theory described in Rosenbaum and Rubin (1983, 1984) and Imbens (2000) shows that the estimated propensity score in step 1 has the correct balancing properties (stated in paragraph 3 of the Introduction) if, for each three-way stratification as shown in Table A2, the distribution of covariates

Table A1: Propensity Scores of Patients i Enrolling in Each of 20 Physician Groups

					The i*th Patient*						
The jth group	1	2	3	4	5	6	7	. . .	i	. . .	2,515
1	$e_{1,1}$	$e_{2,1}$	$e_{3,1}$	$e_{4,1}$	$e_{5,1}$	$e_{6,1}$	$e_{7,1}$	. . .	·	. . .	$e_{2,515,1}$
2	$e_{1,2}$	$e_{2,2}$	$e_{3,2}$	$e_{4,2}$	$e_{5,2}$	$e_{6,2}$	$e_{7,2}$		·		$e_{2,515,2}$
·									·		
j									$e_{i,j}$		
·									·		
20	$e_{1,20}$	$e_{2,20}$	$e_{3,20}$	$e_{4,20}$	$e_{5,20}$	$e_{6,20}$	$e_{7,20}$	. . .	·	. . .	$e_{2,515,20}$

Figure A1: Stratification by Five Strata Based on Propensity Quintiles

Patient	*1*	*2*	*3*	*4*	*5*	*6*	*7*	…	*i*	…	*2,515*
Propensity for the *j* th group	$e_{1,j}$	$e_{2,j}$	$e_{3,j}$	$e_{4,j}$	$e_{5,j}$	$e_{6,j}$	$e_{7,j}$	…	$e_{i,j}$	…	$e_{2,515,j}$

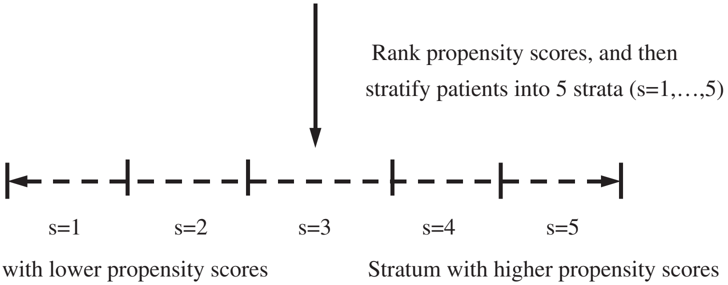
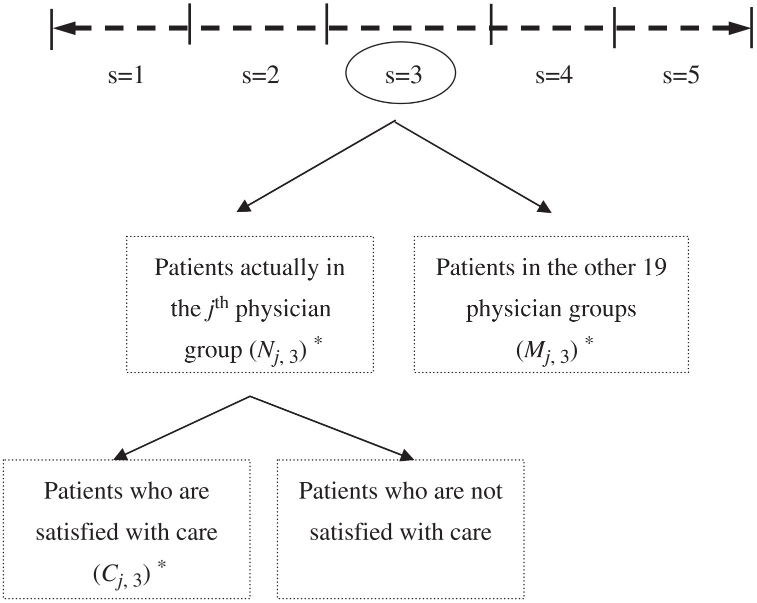

Rank propensity scores, and then stratify patients into 5 strata (s=1,…,5)
s=1
s=2
s=3
s=4
s=5
Stratum with lower propensity scores
Stratum with higher propensity scores

Figure A2: Example of Stratification by Group to Which They Belonged and by Satisfaction with Care


s=1
s=2
s=3
s=4
s=5
Patients actually in the j^{th} physician group ($N_{j,3}$) *
Patients in the other 19 physician groups ($M_{j,3}$) *
Patients who are satisfied with care ($C_{j,3}$) *
Patients who are not satisfied with care

* The same notation as in the Table A2

Table A2: Comparing the jth Physician Group versus Other 19 Physician Groups

	Patients Actually in the j*th Physician Group* (j = *1, . . ., 20)*		*Patients Actually in the Other 19 Physician Groups (Exclude the* j*th Group)*	*Patients in 20 Physician Groups*
Propensity Stratum (s) (1)	*Patients Who Are Satisfied with Care in Strata (s)* (2)	*Total Patients in Strata (s)* (3)	*Total Patients in Strata (s)* (4)	*Total Patients in Strata (s)* (5)
1	$C_{j,1}$	$N_{j,1}$	$M_{j,1}$	$N_{j,1}+M_{j,1}$
2	$C_{j,2}$	$N_{j,2}$	$M_{j,2}$	$N_{j,2}+M_{j\ 2}$
3	$C_{j,3}$	$N_{j,3}$	$M_{j,3}$	$N_{j,3}+M_{j,3}$
4	$C_{j,4}$	$N_{j,4}$	$M_{j,4}$	$N_{j,4}+M_{j,4}$
5	$C_{j,5}$	$N_{j,5}$	$M_{j,5}$	$N_{j,5}+M_{j,5}$
Overall	$\Sigma\ C_{j,s}$	$\Sigma\ N_{j,s}$	$\Sigma\ M_{j,s}$	$\Sigma\ (N_{j,s}+M_{j.s})$

was the same for patients actually in the jth group (column 3) and in the other 19 groups (column 4) within strata of the propensity score. We therefore validated the estimation in step 1 by testing equality (or balance), for each of the physician groups from 20 Tables (not shown), of the distribution for each covariate between columns 3 and 4 within the propensity strata. For these diagnostics, we used two-way ANOVA and logistic regression.

If the variables were not well balanced, as judged by comparison with the expected imbalance merely because of chance, interaction terms of that variable with other variables were estimated into a logistic model along with all previous variables, and a new propensity score was calculated (Rosenbaum and Rubin 1984). As our covariates were well balanced with the original propensity score, this latter step was not necessary in our case.

Step 4: Estimation of Overall Risk-Adjusted Proportion of Satisfied Patients

We estimate the proportion $p_{j,s}$ of patients who would be satisfied with asthma care and who would belong in stratum s if all patients had been enrolled in the jth physician group ($j = 1, \ldots, 20$), by:

$$P_{j,s} = C_{j,\,s}/N_{j,\,s}$$

For each physician group, we estimate the overall risk-adjusted proportion p_j of satisfied patients using the weighted average by combining across the five strata, by:

$$P_j^+ = \sum \left(P_{j,s} * W_{j,s}\right), \text{ with weights}$$
$$W_{j,s} = \left(N_{j,s} + M_{j,s}\right) / \sum \left(N_{j,s} + M_{j,s}\right)$$

Step 5: Estimation of the Relative Performance of Each Physician Group Using Shrinkage Techniques

We estimated the log odds, $y_j = \log\left[p_j/(1 - p_j)\right]$, of satisfaction with asthma care for each physician group by $Y_j = \log\left[P_j^+ / \left(1 - P_j^+\right)\right]$, using the estimated overall risk-adjusted proportions P_j^+ obtained in step 4. Variances estimates V_j of Y_j were obtained by the delta method. To address regression-to-the-mean associated with comparing multiple physician groups, we then adjusted these preliminary log odds estimates of physician group performance towards the grand mean using the shrinkage method by Morris (1983).

This method, a generalization of James and Stein's (1961) estimation, is used to improve the performance of approximately independent normal and unbiased estimators ($Y_j, j = 1, \ldots, 20$) of corresponding parameters—here the underlying true log odds, y_j, of patient satisfaction that can be offered by physician group j, $j = 1, \ldots, 20$. For each parameter y_j, the method produces a new estimator, Y_j^f, that is a weighted combination of the original estimator Y_j for that parameter and of a weighted average Y^f, of all original estimators. The weights are such that the new estimator is increasingly shrunk away from the original one and towards that average of all estimators when the variance V_j within estimator Y_j is large relative to the variance across estimators.

Specifically, the weights F_j and the weighted average estimator Y^f are defined by the relations

$$F_j = 1/(V_j + A),$$

$$Y^f = \frac{\sum_{j=1}^{20} F_j Y_j}{\sum_{j=1}^{20} F_j} \quad \text{and where} \quad A = \max\left(\frac{20\sum_{j=1}^{20} F_j\left\{(Y_j - Y^f)^2 - V_j\right\}}{19\sum_{j=1}^{20} F_j}, 0\right)$$

The solution for Y^f and A from the above equations is obtained iteratively. Then, the shrinkage estimator is given by

$$Y_j^f = (1 - B_j) Y_j + B_j Y^f \quad \text{where} \quad B_j = (17/19) V_j / (V_j + A)$$

is the shrinkage factor.

Finally, we estimated the overall risk-adjusted OR of physician group performance in comparing the jth physician group ($j = 2, \ldots, 20$) versus physician group 1 by exponentiating the difference of the corresponding log odds. Because the estimates from Morris's approach are dependent across different groups, the standard errors for these and for the final estimates of the ORs were calculated by simulation. This was done by first simulating a large number of sets of 20 independent normal draws, centered at the means Y_j^f that were estimated from the shrinkage estimator and with variances V_j. Using each set as data, we obtained estimates of the means from which they were drawn using Morris's method, and then we computed the sample variance–covariance matrix of these estimates across the sets.

References

Alter, D. A., P. C. Austin, C. D. Naylor, and J. V. Tu. 2002. "Factoring Socioeconomic Status into Cardiac Performance Profiling for Hospitals: Does It Matter?" *Medical Care* 40 (1): 60–7.

Aron, D. C., D. L. Harper, L. B. Shepardson, and G. E. Rosenthal. 1998. "Impact of Risk-Adjusting Cesarean Delivery Rates When Reporting Hospital Performance." *Journal of the American Medical Association* 279 (24): 1968–72.

Bodenheimer, T. 1999. "The American Health Care System—The Movement for Improved Quality in Health Care." *New England Journal of Medicine* 340 (6): 488–92.

Braitman, L. E., and P. R. Rosenbaum. 2002. "Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores." *Annals of Internal Medicine* 137 (8): 693–5.

Christiansen, C. L., and C. N. Morris. 1997. "Improving the Statistical Approach to Health Care Provider Profiling." *Annals of Internal Medicine* 127 (8, Part 2): 764–8.

Cochran, W. G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24 (2): 295–313.

Connors, A. F. Jr., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell Jr., D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson Jr., H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus. 1996. "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients. Support Investigators." *Journal of the American Medical Association* 276 (11): 889–97.

D'Agostino, R. B. Jr. 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* 17 (19): 2265–81.

Dehejia, R. H., and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

DeLong, E. R., E. D. Peterson, D. M. DeLong, L. H. Muhlbaier, S. Hackett, and D. B. Mark. 1997. "Comparing Risk-Adjustment Methods for Provider Profiling." *Statistics in Medicine* 16 (23): 2645–64.

Diette, G. B., A. W. Wu, E. A. Skinner, L. Markson, R. D. Clark, R. C. McDonald, J. P. Healy Jr., M. Huber, and D. M. Steinwachs. 1999. "Treatment Patterns among Adult Patients with Asthma: Factors Associated with Overuse of Inhaled Beta-Agonists and Underuse of Inhaled Corticosteroids." *Archives of Internal Medicine* 159 (22): 2697–704.

DiGiuseppe, D. L., D. C. Aron, S. M. Payne, R. J. Snow, L. Dierker, and G. E. Rosenthal. 2001. "Risk Adjusting Cesarean Delivery Rates: A Comparison of Hospital Profiles Based on Medical Record and Birth Certificate Data." *Health Services Research* 36 (6): 959–77.

Enthoven, A. C. 1993. "The History and Principles of Managed Competition." *Health Affairs (Millwood)* 12 (suppl): 24–48.

Everson, P. J., and C. N. Morris. 1993. *Splus Software: Hierarchical Normal Regression Model.* Boston: Harvard University.

Fayers, P. M., and D. Machin. 2000. *Quality of Life: Assessment, Analysis and Interpretation.* Chichester, UK: John Wiley & Sons, Ltd.

Fowler, F. J. Jr., P. M. Gallagher, V. L. Stringfellow, A. M. Zaslavsky, J. W. Thompson, and P. D. Cleary. 2002. "Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members." *Medical Care* 40 (3): 190–200.

Gum, P. A., M. Thamilarasan, J. Watanabe, E. H. Blackstone, and M. S. Lauer. 2001. "Aspirin Use and All-Cause Mortality among Patients Being Evaluated for

Known or Suspected Coronary Artery Disease: A Propensity Analysis." *Journal of the American Medical Association* 286 (10): 1187–94.

Hirano, K., and G. W. Imbens. 2002. *Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.* New York: National Bureau of Economic Research.

Iezzoni, L. I. 1997. *Risk Adjustment for Measuring Healthcare Outcomes.* Chicago: Health Administration Press.

Imbens, G. W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87 (3): 706–10.

James, W., and C. Stein. 1961. *Estimation with Quadratic Loss.* Berkeley, CA: University of California Press.

Kahn, K. L., M. L. Pearson, E. R. Harrison, K. A. Desmond, W. H. Rogers, L. V. Rubenstein, R. H. Brook, and E. B. Keeler. 1994. "Health Care for Black and Poor Hospitalized Medicare Patients." *Journal of the American Medical Association* 271 (15): 1169–74.

Katon, W., C. M. Rutter, E. Lin, G. Simon, M. Von Korff, T. Bush, E. Walker, and E. Ludman. 2000. "Are There Detectable Differences in Quality of Care or Outcome of Depression across Primary Care Providers?" *Medical Care* 38 (6): 552–61.

Keating, N. L., J. C. Weeks, M. B. Landrum, C. Borbas, and E. Guadagnoli. 2001. "Discussion of Treatment Options for Early-Stage Breast Cancer: Effect of Provider Specialty on Type of Surgery and Satisfaction." *Medical Care* 39 (7): 681–91.

LaLonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experiment Data." *The American Economic Review* 76 (4): 604–20.

Masland, M., A. W. Wu, G. B. Diette, F. Dominici, and E. A. Skinner. 2000. "The 1998 Asthma Outcomes Survey." San Francisco: Pacific Business Group on Health.

Mojtabai, R., and J. G. Zivin. 2003. "Effectiveness and Cost-Effectiveness of Four Treatment Modalities for Substance Disorders: A Propensity Score Analysis." *Health Services Research* 38 (1, Part 1): 233–59.

Morris, C. N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78 (381): 47–55.

Rosenbaum, P. R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 1024–32.

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

———. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24.

———. 1985. "The Bias Due to Incomplete Matching." *Biometrics* 41 (1): 103–16.

Rubin, D. B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8, part 2): 757–63.

Shahian, D. M., S. L. Normand, D. F. Torchiana, S. M. Lewis, J. O. Pastore, R. E. Kuntz, and P. I. Dreyer. 2001. "Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique." *Annals of Thoracic Surgery* 72 (6): 2155–68.

Shwartz, M., R. Saitz, K. Mulvey, and P. Brannigan. 1999. "The Value of Acupuncture Detoxification Programs in a Substance Abuse Treatment System." *Journal of Substance Abuse Treatment* 17 (4): 305–12.

Streiner, D. L., and G. R. Norman. 1995. "Health Measurement Scales a Practical Guide to Their Development and Use." Oxford, UK: Oxford University Press.

Steinwachs, D. M., A. W. Wu, and E. A. Skinner. 1994. "How Will Outcomes Management Work?" *Health Affairs (Millwood)* 13 (4): 153–62.

Sullivan, L. M., K. A. Dukes, and E. Losina. 1999. "Tutorial in Biostatistics. An Introduction to Hierarchical Linear Modelling." *Statistics in Medicine* 18 (7): 855–88.

Ware, J. E. Jr. 1997. *Sf-36 Physical & Mental Health Summary Scales: A User's Manual.* Boston: Quality Metric.

Welch, H. G., W. C. Black, and E. S. Fisher. 1995. "Case-Mix Adjustment: Making Bad Apples Look Good." *Journal of the American Medical Association* 273 (10): 772–3.